ORIE 4741 mid-term report

Nanqing Dong (nd367), Ziyi Chen (zc286)

Introduction

We aim at predicting whether an inpatient is likely to survive, which helps recommending customized hospitals and treatments for him (her) to lower mortality risk. Such a prediction can be formulated into a binary classification problem, in which we train the classfier using the Statewide Planning and Research Cooperative System (SPARCS) Hospital Inpatient Discharges dataset [1]. This dataset has the discharge records with 39 variables of 2544731 patients. Among these variables, the "inpatient disposition" tells us whether an inpatient survived upon discharge. Detail of these variables is shown in Table 1.

Data Preprocessing

At first, we deleted irrelevant variables by common sense and kept 16 predictors as well as the variable called "inpatient disposition". Then, we transformed the "inpatient disposition" into the target variable called "die", indicating whether a patient passed away ("die=1") or not ("die=0").

Moreover, we delete the 100130 patients with missing values, including global missing value "na", and those specific to variables like "Ethnicity=Unknown", "Gender=U", etc.

Preliminary observation on the relevancy of each predictor

Since all the predictors are categorical, we measure the relevancy of a predictor X_j by the variation of the death rate with each value k of X_j. The death rate for $X_j = k$ is defined as the percentage of the dead inpatients in the inpatients whose $X_j = k$. For each variable, we roughly measure its relevancy by the standard deviation of these death rates, as shown in Table 1.

It can be seen from Table 1 that the standard deviation of death rate for the variable "APR Risk of Mortality", the diagnosed disease and its severity are among the highest, and that of the inpatients' gender and admitted day of the week are among the lowest, which also makes sense. Fortunately, the hospital ("facility id") and treatment ("CCS Procedure Code", "APR MDC Code") we concern have comparable standard deviation to the highest one.

To go further into the relevancy of each predictor, we plot barcharts of these death rates for "APR MDC Code" and "Admit Day of Week", in Figure 1. Figure 1 (a) shows a large variation of death rate with "APR MDC Code" (category of diagnosed disease). Among the diseases, "Infectious and Parasitic Diseases, Systemic or Unspecified Sites" (18) has the highest death rate, whereas "Pregnancy, Childbirth and the Puerperium" (14) has the lowest death rate, which also fits common sense. Figure 1 (b) shows very slight effect of inpatient's admitted day on the death rate. As it approaches the weekend, the death rate increases and reach the peak during the weekend, and dramatically falls to the lowest on Monday. In fact, this is reasonable since most staffs tend to be more responsible on weekday than on weekend.

Feature selection by mutual information

Feature selection is an important method to avoid overfitting. Since the target and predictors are all categorical, we adopt a simple feature selection method based on the mutual information (MI) between each predictor and the target variable, a commonly used measure of relevancy between the two [3]. We compare these MI's and list them in decreasing order in Table 1. As shown in Table 1, MI is significantly correlated with standard deviation of death rate. We select the 10 predictors with the highest MI's from "APR Risk of Mortality" to "Emergency Department Indicator".

Naive Bayes Classification

After feature selection, we adopt Naive Bayes classifer [3], a simple but popular classification method for categorical predictors.

To test the performance, we randomly partition the dataset into training dataset with 80% of the samples and test dataset with 20% of the samples. To avoid overfitting, we train the classifier with various smoothing parameter λ using 5-fold cross-validation on the training dataset. Since there is a significant imbalance between the positive group (52721 samples whose "die=1") and negative group (2391880 samples whose "die=0"), we adopt F1 score instead of the proportion of misclassification error as the measure of performance. Finally, we list the F1 score on the test dataset and the mean F1 score on the 5 validation sets in Table 2, for $\lambda \in \{0, 0.1, 1, 2, 3, 5, 10, 100\}$ respectively.

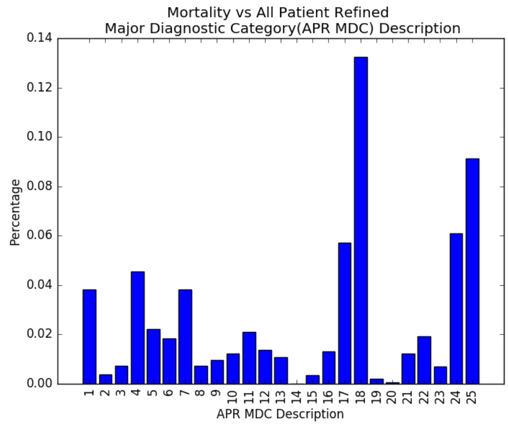
It can be seen from Table 2 that all these F1 scores are all slightly more than 0.30, far from the full mark 1.0, thus it requires more work later to improve the performance. In addition, the F1 scores on both the validation data and the test data keep slightly increasing with λ, which means it is likely to get more accurate prediction with larger λ.

Future work

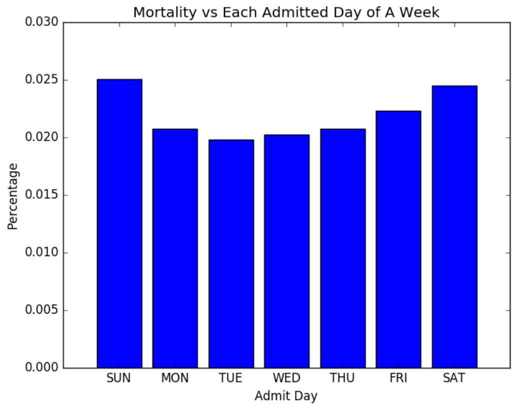
1. Delete highly correlated predictors.
2. Delete samples: Some predictors have over 200 values, whereas some values are not representative enough due to limited samples, which could be deleted.
3. We will try various ways to deal with imbalance between the 2 classes, such as bootstrap.
4. Try more classifiers, and then use cross validation to choose the best classifier.
5. Use the best classifier to recommend customized hospitals and treatments for a specific new patient that have small predicted death rate.

Variable name	Explanation	The number of values	Standard deviation of death rate	Mutual information with target variable
APR Risk of Mortality	The likelihood of mortality estimated via diagnosis, encoded as: 1=Minor, 2=Moderate, 3=Major, 4=Extreme. Note: This estimation is based on severity of disease but ignores factors like hospitals and treatments. [2] We seek to improve accuracy by adding these factors.	4	0.1033	0.03674
APR Severity of Illness Code	Severity of illness diagnosed by the APR-DRG system, encoded as: 1=Minor, 2=Moderate, 3=Major, 4=Extreme.	4	0.0851	0.03036
APR DRG Code	Diseases diagnosed by the APR-DRG system, encoded by integers.	313	0.0867	0.02812
CCS Diagnosis Code	Diseases diagnosed by the CCS system, encoded by integers.	260	0.0556	0.02341
CCS Procedure Code	Treatment adopted categorized by the CCS system, encoded by integers.	230	0.0440	0.02052
APR MDC Code	Treatment adopted categorized by the APR-MDC system	25	0.0308	0.01300
Age group	The inpatient's age group: "0 to 17", "18 to 29", "30 to 49", "50 to 69", "70 or Older"	5	0.0184	0.00920
Facility id	Hospital code	220	0.0584	0.00546
Type of admission	The method in which was the inpatient admitted to the hospital	6	0.0141	0.00232
Emergency Department Indicator	Whether the revenue record contained an Emergency Department revenue code of 045X	2	0.0069	0.00116
Hospital county	Which county is the hospital in?	56	0.0081	0.00052
APR Medical Surgical Description	Categorize the treatment into {surgical, medical} by the APR-DRG system	2	0.0050	0.00046
Race	The race of the inpatient	3	0.0040	0.00044
Ethnicity	The ethnicity of the inpatient	2	0.0040	0.00020
Gender	The gender of the inpatient	2	0.0024	0.00014
Admit Day of Week	The day of week in which the inpatient was admitted	7	0.0019	0.00007

Table 1: The 16 categorical predictors [1]



(a) “APR MDC Code”



(b) “Admit Day of Week”

Figure 1: The death rate for each value

Smoothing parameter λ	The mean F1 score on the 5 validation datasets	The F1 score on the test dataset
0	0.3048	0.3086
0.1	0.3051	0.3088
1	0.3054	0.3089
2	0.3057	0.3090
3	0.3060	0.3093
5	0.3064	0.3097
10	0.3067	0.3102
100	0.3071	0.3104

Table 2: The performance of naive Bayes classifier

References

[1] Statewide Planning and Research Cooperative System (SPARCS) Hospital Inpatient Discharges dataset:
https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De
-Identified/u4ud-w55t

[2] Baram, Daniel, et al. ”Use of the All Patient Refined-Diagnosis Related Group (APR-DRG) risk of mortality score as a severity adjustor in the medical ICU.” Clinical Medicine Insights. Circulatory, Respiratory and Pulmonary Medicine 2 (2008): 20.

[3] Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.